Die Suchmaschine SENTRAX. Grundlagen und Anwendungen dieser Neuentwicklung

Hans-Joachim Bentz

Universität Hildesheim
Institut für Mathematik und Angewandte Informatik
Marienburger Platz 22
31141 Hildesheim
bentz@cs.uni-hildesheim.de

Zusammenfassung

Es wird eine intelligente Suchmaschine für den bequemen Zugriff auf strukturierte und unstrukturierte Informationen vorgestellt. Grundlage bilden 4 verschiedene Ähnlichkeitsmaße auf den Datensorten in der Datenbasis gemäß den jeweiligen Aufgaben: Schreibweisentolerante Suche, Kontextähnliche Suche, Zugriff auf Dokumententreffer, Doublettensuche.

Abstract:

This article introduces an automatic "essence-extractor-engine" which works both on structured and inhomogenous document collections and supports interactive searching.

1 Einleitung

An einem vernetzten Lern- bzw. Arbeitsplatz, wo sowohl unterschiedliche Aufgaben, Anforderungen und Prozesse wirken als auch verschiedenartige Arbeitsmaterialien und Datenbasen in den Zugriff gestellt werden, ist die "ad hoc-Suche" nur selten erfolgreich. Unter anderem liegt es an den Schwächen herkömmlicher Suchtechnik, welche meist auf Matching-Algorithmen auf sogenannten "invertierten Listen" beruht. Diese Listen enthalten praktisch alle Wörter aus den Texten, wegen des schnellen Zugriffs alphabetisch sortiert, zusammen mit den Herkunftskoordinaten. Weicht der Suchbegriff auch nur wenig vom Listenwort ab, insbesondere im ersten Wortteil, wie zum Beispiel Apartment und Appartement, dann geht die Suche fehl oder bleibt unvollständig. Man hat mehrere Abhilfen versucht (wie etwa eine Rückwärtstrunkierung, wohinter sich eine invertierte Liste aller Wörter von hinten gelesen verbirgt; oder eine Reduktion der Wörter auf ihren Wortstamm samt Zerlegung in Wortbestandteile, wie Arbeit-s-Material-ien; oder auch zusätzliche Verweise etc.), jedoch greift keine davon entscheidend durch. Die SENTRAX Engine verfolgt einen anderen Ansatz: Texte werden als abstrakte Sequenzen von Strings über einem festzulegenden Alphabet interpretiert und darauf dann Mustererkennung betrieben. Je nach Aufgabenstellung setzt eine passende Routine ein, die mit einem auf die Problemstellung zugeschnittenen Ähnlichkeitsmaß Muster sucht und erkennt. Somit

werden (auf Stringebene) Tippfehler bzw. Schreibvarianten toleriert oder (auf Wort- bzw. Satzebene) bedeutungsverwandte Begriffe zugelassen. Es werden vier Aufgabenstellungen unterschieden und unterstützt:

1) *lexico*: die Suche nach ähnlichen Zeichenketten,
2) *context*: die Suche nach bedeutungsverwandten bzw. im gleichen Kontext vorkommenden Begriffen,
3) *treffer*: das auf Wortebene tolerante Holen von Dokumenten mit den Suchbegriffen,
4) *similardoc*: das auf Dokumentebene tolerante Präsentieren von Doubletten oder Fastdoubletten.

SENTRAX steht für Essence Extractor Engine. Das Hauptmotiv bei der Entwicklung der Suchmaschine war, das Problem der meist unscharfen Konzeptbeschreibung in den Griff zu bekommen. Für den Benutzer ist es immer schwierig, bei seiner Anfrageformulierung geeignete Begriffe zu finden, die einerseits sein (vages) Informationsbedürfnis wiedergeben und andererseits die geeignete Grundlage für die Suche im System bieten. Wegen der Vielzahl an Möglichkeiten, ein und denselben Vorgang oder Sachverhalt zu beschreiben, ist es nicht leicht, mit einer begrenzten Anzahl von Suchbegriffen die Weite eines Konzepts zufriedenstellend abzudecken. Um hierbei zu helfen befindet sich im SENTRAX-Container eine (maschinelle) Zusammenstellung solcher Begriffscluster aus den Texten, die geeignet sind bestimmte Konzepte zu verkörpern. Über eine mehrstufige Kookkurrenzanalyse (basierend auf den Untersuchungen von Wettler et.al [1995] und der Arbeit von Ackermann [2000]) werden „Wortwolken“ gebildet, die die „Essenz der Texte“ repräsentieren sollen. Die mehrdimensionale Wortwolke wird dann zweidimensional auf den Bildschirm projiziert.

2 Typische Hindernisse bei herkömmlicher Technologie

In der nachfolgenden Übersicht sind ein paar Beispiele von Wörtern und Wortkonstruktionen zusammengestellt, die Probleme bei der computergestützten Suche mit sich bringen. Alle Beispiele entstammen realen Dokumenten, sind also in der Praxis vorgekommen. Die Gruppierung ist willkürlich vorgenommen, teilweise überschneiden sich die Typen. Sie sollten sich ohne größere Erläuterung verstehen lassen. Lediglich das Beispiel "method-rnethod" bedarf eines Kommentars: in einer PDF-Datei stand tatsächlich "r-n" anstatt m (OCR Fehlgriff?), weshalb das lesende Auge hier keinen "Fehler" sah, die SENTRAX Engine jedoch einen solchen auswies. Bei genauerer Inspektion löste sich das wie beschrieben auf. Die Phänomene, ob nun tückisch oder simpel, sind natürlich nicht nur auf deutsche Texte beschränkt, vergleichbare Fälle gibt es auch in anderen Sprachen.

Suchworte sind nicht exakt so im Text enthalten, wie erwartet
(Herrmann–Hermann, Ullrich-Ulrich, Detlef–Detlev, Maßstab–Massstab, Notfallmaßnahme–Notfallschutzmaßnahme, Uranbergbau–Uranerzbergbau)

Suchworte haben zulässige oder tolerierte Schreibvarianten
(Foto – Photo, Fahrkostenerstattung – Fahrtkostenerstattung, grey – gray, Potenzial – Potential, Apartment-Appartement, Bundesforschungsminister – Bundesminister für Forschung und Technologie, Numerierung - Nummerierung)

Suchworte sind im Singular, im Dokument aber im Plural oder umgekehrt

(Visum – Visa, Universum - Universen)

Suchworte sind zwar korrekt eingegeben, im Dokument aber fehlerhaft
(Archivierung-Archiverung, Libyen–Lybien, method-rnethod)

Eingabe ist falsch geschrieben (Mitterand) **oder hat Tippfehler**
(Grundwasserstömung, refernce)

Anstelle des Suchworts steht „leider" ein bedeutungsverwandter Begriff im Text
(Fusion–Zusammenschluß-Verschmelzung, Wegbeschreibung–Anfahrtplan, Firmenpleiten–Konkurse–Insolvenzen)

3 Vom Suchen zum Finden durch vier Optionen

In diesem Abschnitt wird kurz die "normale" Arbeitsweise mit den SENTRAX-Funktionen beschrieben. Dazu stelle man sich vor, einen Datenbestand zu haben, der aus recht unstrukturiertem Material bestehen mag, zum Beispiel aus x1000 Office-Dokumenten (HTML, TXT, WORD, PDF, PPT, EXCEL). Man kann sie im File-System liegen haben oder auch in einer Datenbank. Die SENTRAX-Engine geht in der Standardanwendung alle Dokumente einmal durch und "liest" sie zum Zweck der Erstellung eines "Containers", der einem Index entspricht. Aus praktischen Gründen wird von jedem Dokument eine HTML-Version hergestellt. Obgleich dadurch manchmal gewisse Formatierungen und Objekte verloren gehen, hat man zum Ausgleich erstens einen sehr schnellen Zugriff auf den Text - auch über das Inter- oder Intranet- (man braucht also nicht eigens eine Applikation, z.B. ein WORD-Programm, zu öffnen) und zweitens die Möglichkeit die Fundstellen anzuspringen und hervorzuheben (z.B. durch eine Highlight-Funktion). Wer einmal versucht hat, durch ein 10 MByte großes PDF-Dokument zu blättern, weiß die beschriebene Option zu schätzen! Nachdem eine später als Trefferkandidat angesehene HTML-Version positiv geprüft wurde, kann natürlich mit einem Klick das zugehörige Original aufgerufen und ggf. bearbeitet werden (vgl. unten: (5) Ansichtsoptionen).

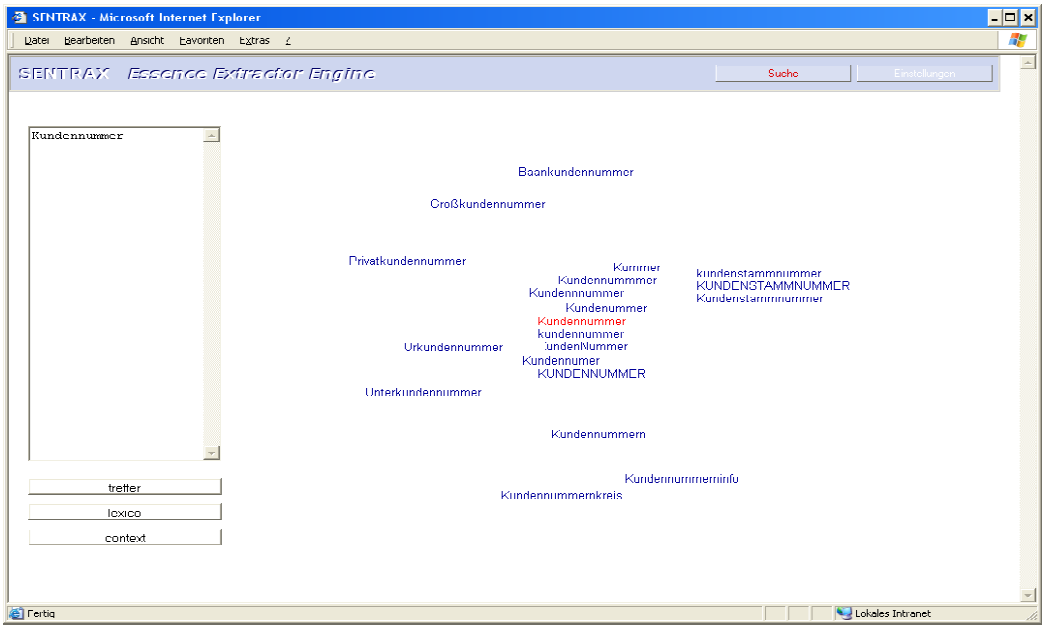

Abb. 1: SENTRAX Bildschirm nach Eingabe: "Kundennummer" und Aktivierung der *lexico*-Funktion

(1) Die *lexico*-Funktion

Bei der ersten Eingabe empfiehlt sich, das Suchwort der *lexico*-Funktion zu übergeben. Man erhält dann eine (manuell einstellbare) Anzahl von Begriffen, die String-ähnlich zur Eingabe sind und alle im Dokumentenbestand vorkommen. Sie werden in einer Map zweidimensional dargestellt, wobei sich die Kandidaten um das ins Zentrum platzierte Suchwort, das farblich hervorgehoben ist, gruppieren. Je näher dran, desto ähnlicher. Diese Ausgabe kann auch (manuell einstellbar) in Form einer Liste gegeben werden. Besonders hilfreich ist die *lexico*-Funktion in dem Fall, wo das Eingabewort nicht in der Datenbasis vorkommt. Man sieht diesen Umstand sofort, hat aber automatisch die Menge der Angebote von schreibweisenähnlichen, aus denen man sich dann eines oder mehrere für die weiteren Suchschritte wählen kann. Überdies erkennt man stets vorkommende Tippfehler, Schreibvarianten oder Begriffsvariationen.

So erkennt man in dem Beispiel in Abbildung 1 erstens: das Eingabewort "Kundennummer" gibt es in den Daten, zweitens: es kommt außerdem noch mit Tippfehler vor (Kundenummer), drittens: es gibt diverse Kompositionen, die für die Suche interessant oder hilfreich sein können (Kundenstammnummer, Privatkundennummer, Kundennummernkreis etc.).

Durch Anklicken eines anderen Worts wird dessen Umgebung auf den Schirm geholt, durch nochmaliges Anklicken (toggeln) eines bereits gefärbten Worts wird dieses wieder ausgeschaltet. Durch zusätzliche Eingabe eines Wortes ins Eingabefenster kommt dessen Umgebung dazu. Nachdem man sich für ein oder mehrere Wörter entschieden hat, die für die weiteren Schritte verwendet werden sollen, kann man entweder die *context*-Funktion oder die *treffer*-Funktion aktivieren.

(2) Die *context*-Funktion

Nach Eingabe von Begriffen oder Auswahl aus der lexico-Map kann die *context*-Funktion aktiviert werden. Sie projiziert eine (im Begriffsumfang manuell einstellbare) Sammlung von Wörtern aus dem Container auf den Bildschirm. Diese Sammlung beinhaltet solche Begriffe, die prädominant im Kontext mit den Suchwörtern der Eingabe vorkommen. Dabei sind auch höhere Ordnungen mitberechnet, weshalb es passieren kann, dass zwei Begriffe als nahe eingestuft und gezeigt sind, obwohl sie nie zusammen in demselben Dokument vorkommen. Auch hier kann man wieder weitere vom Bildschirm dazuklicken oder wegklicken. Damit wird eine interaktive, zielgerichtete Suche möglich. Insbesondere zeigt sich bei den meisten Nutzern, daß die angebotenen Wortgruppen Teilkonzepte oder verbundene Konzepte aufrufen, also wie ein "passives Gedächtnis" fungieren. Somit leistet die SENTRAX Technologie einen wichtigen Beitrag zur vollautomatischen Erschließung von Sinnstrukturen im Datenbestand.

Ein treffendes Beispiel für die diversen Eigenschaften zeigt die Abbildung 2. Der Bildschirm ist einer Arbeit von Kummer [2006] entnommen, wo Trefferlisten herkömmlicher Suchmaschinen mit Hilfe der SENTRAX-Technologie analysiert werden. Im konkreten Beispiel wurde das Wort "Todsünden" in eine der Internetsuchmaschinen eingegeben und ein Satz der angebotenen Trefferseiten dann in einem SENTRAX-Container erfasst. Das gleiche Suchwort lieferte nun (über die *context*-Funktion) den gezeigten

Screen. Man kann mit einem Schlag alle Todsünden erkennen als auch weitere Begriffe, die auf den fraglichen Seiten im Zusammenhang mit dem Suchkonzept stehen.

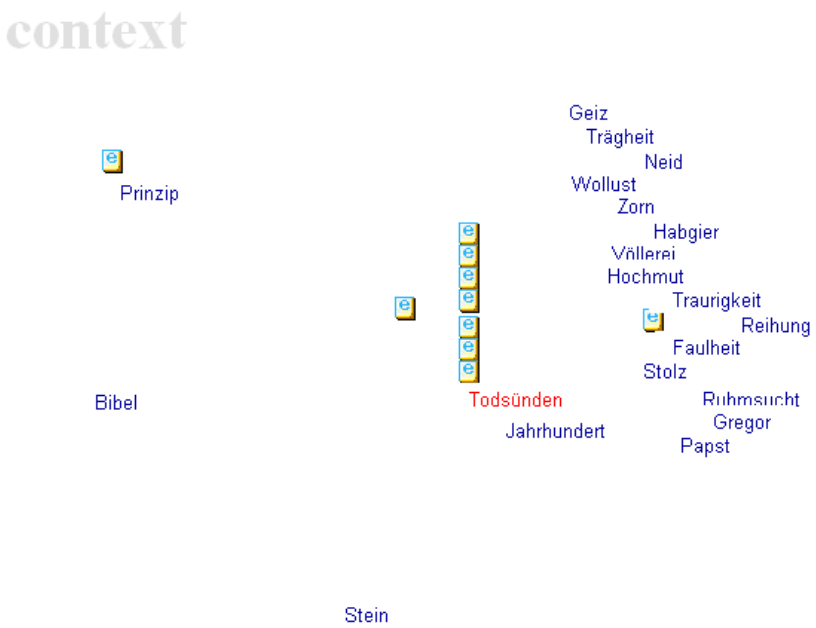

Abb. 2: SENTRAX Bildschirm nach Eingabe: "Todsünden" mit der *context*-Funktion (aus Kummer [2006])

Im Beispiel der Abb. 3 wurden zwei Wörter -"kreative Selbstpräsentation"- in eine der Internetsuchmaschinen eingegeben und wieder ein Anfangsstück der gemeldeten Treffer-

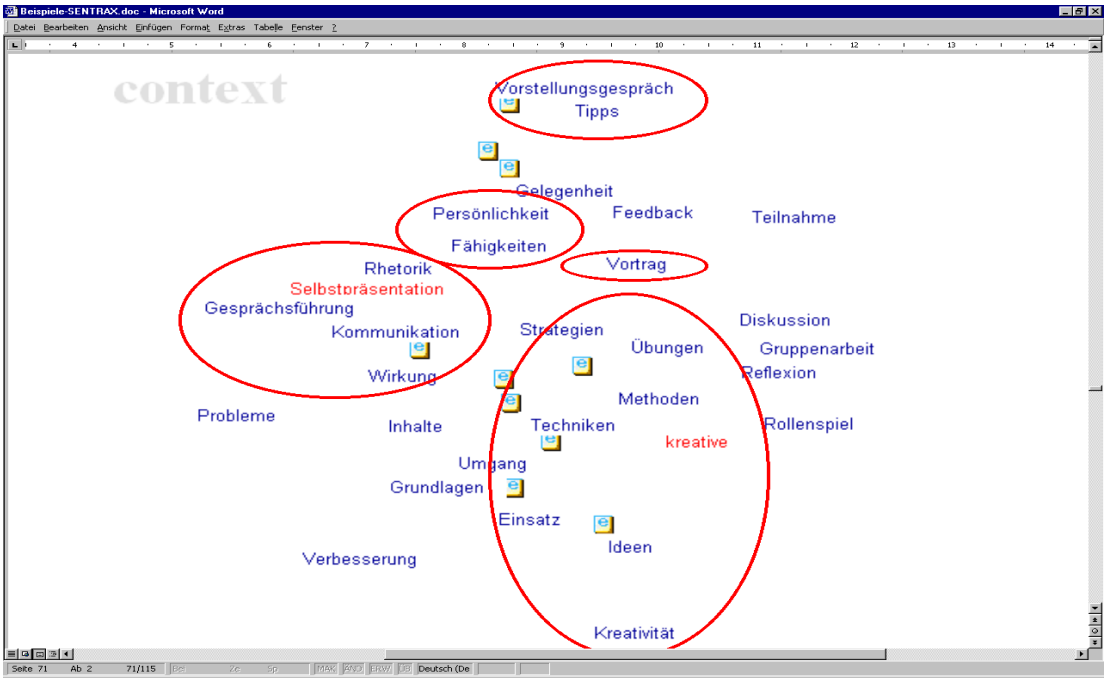

Abb. 3: SENTRAX *context*-Screen nach Eingabe: "kreative Selbstpräsentation" (aus Kummer [2006])

seiten mit der SENTRAX verarbeitet. Man kann ganz gut mehrere Teilcluster erkennen, die unterschiedliche Aspekte der "kreativen Selbstpräsentation" darstellen. Die Markierungen sind hier manuell zum Zweck der besseren Illustration ergänzt.

(3) Die *treffer*-Funktion

Normalerweise gibt es zu einem Suchwort viele Trefferdokumente, in denen es mit dieser oder jener Bedeutung enthalten ist. Ein Teil davon wird durch entsprechende Symbole bereits auf dem *context*-Screen gezeigt und kann von dort direkt aufgerufen werden. Dieses Vorgehen ist im allgemeinen aber nicht besonders effizient. Man will ja wenn möglich vermeiden, manuell viele Dokumente durchzulesen um zu sehen, ob sie als "Lösung" des Suchproblems in Frage kommen. Vielmehr hat man durch die bequeme Komplettierung des gesuchten Konzepts mittels Hinzuklicken passender Wörter aus der *context*-Wortwolke die Möglichkeit, die Anzahl der in Frage kommenden Dokumente einzuschränken. Je spezifischer nämlich die Beschreibung der Suchidee wird, um so weniger Treffer wird es normalerweise geben. Sobald die Anzahl der Dokument-Icons übersichtlich geworden ist, kann man sich mit der treffer-Funktion spezifischere Informationen zeigen lassen und einen Schnellzugriff auf ein eventuell interessierendes Dokument vorbereiten.

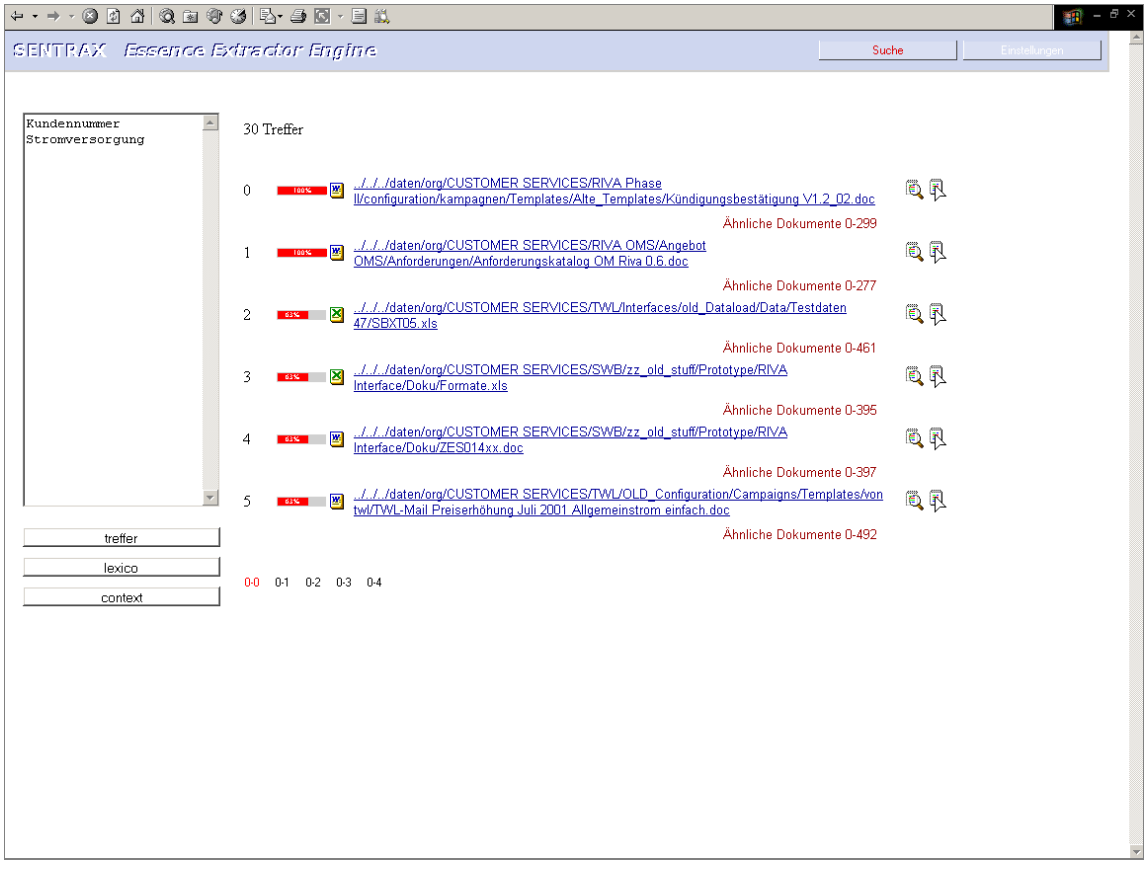

Abb. 4: SENTRAX *treffer*-Screen nach Eingabe: "Kundennummer Stromversorgung"

Im Beispiel der Abbildung 4 sieht man die ersten sechs von 30 gemeldeten Treffern mit Dokumentname und Pfad und weiteren Optionen. Es ist auch einstellbar, die z.B. ersten

300 Zeichen aus dem Text gezeigt zu bekommen und weitere Metainformationen (wie Autor, Dokumentgröße usw.), um sich in diesem Stadium der Suche bequemer orientieren zu können. Dieser Ausgabescreen entspricht in etwa der Situation bei herkömmlicher Internetsuche: Man gibt etwas ein und erhält eine Liste von "Treffern".

In der konkreten Suche der Abbildung 4 gab es zwei Eingabewörter deren Vorkommen im jeweiligen Trefferdokument durch den Füllgrad des roten Balkens angezeigt wird. Ist er voll gefüllt, dann sind alle Eingaben enthalten, ist er nur teilgefüllt, dann fehlt eines usw. Die Ausgabeliste wird standardmäßig nach dem Füllgrad sortiert. Die Symbole rechts in den Zeilen sind Buttons und erlauben die schnelle Ansicht des Dokuments im HTML-Format. Ein Klick auf das Dokument selbst öffnet dieses.

(4) Die Funktion "Ähnliche Dokumente" (*similarDoc*)

Diese Funktion ist die letzte der vier SENTRAX-Suchoptionen. Hier beruht das Ähnlichkeitsmaß auf einer gewichteten Anzahl gemeinsamer (bedeutungsvoller) Wörter. Man aktiviert diese Funktion für ein Dokument, indem man im *treffer*-Screen das einem Dokument zugehörige Feld "Ähnliche Dokumente" anklickt. Es muss nicht das oberste Dokument sein, sondern kann beliebig aus der Trefferliste gewählt werden.

Als Folge wird eine Liste angezeigt, die vom aktivierten Ausgabedokument angeführt wird. Darunter stehen dann mit absteigender Ähnlichkeit die Kandidaten. Der Grad der Ähnlichkeit, den die Engine ermittelt, wird wieder durch die Füllung des Balkens bzw. auch durch den darin eingetragenen %-Wert ausgedrückt.

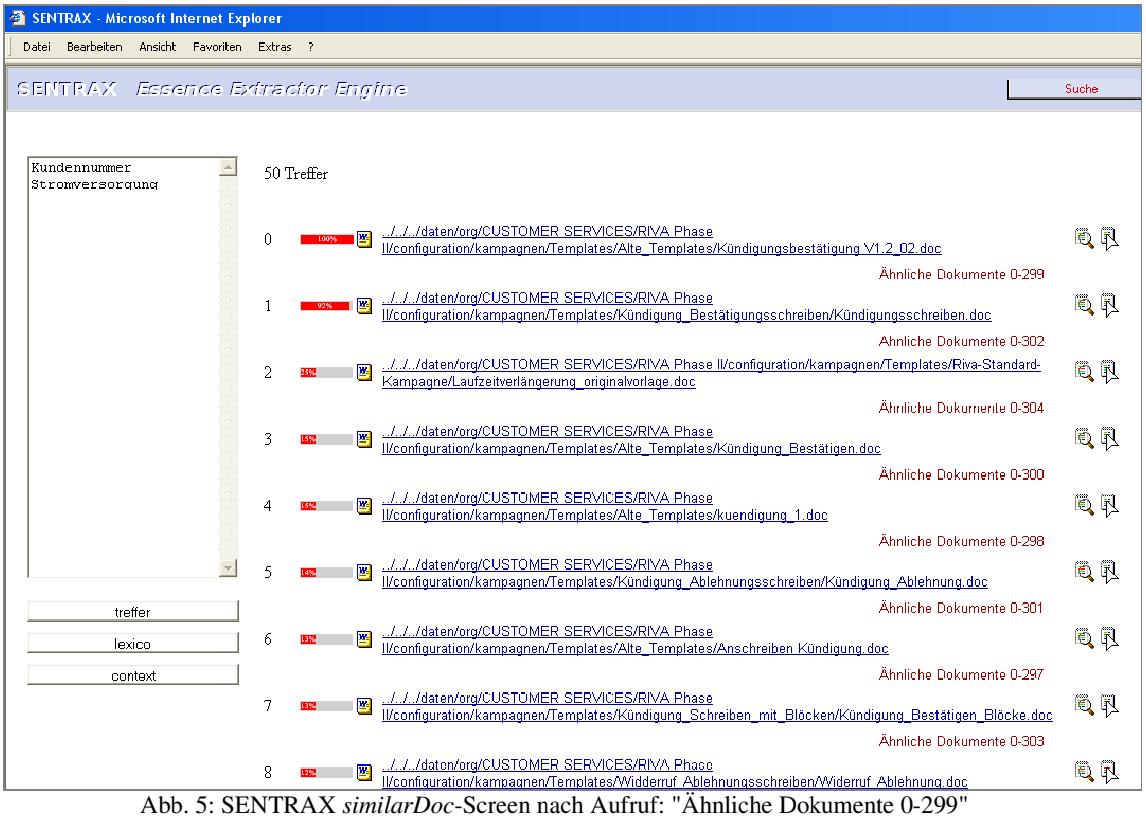

Abb. 5: SENTRAX *similarDoc*-Screen nach Aufruf: "Ähnliche Dokumente 0-299"

Nach unseren Erfahrungen ergeben sich beim intellektuellen Nachprüfen (durch Ansehen der Kandidaten) schon recht große Ähnlichkeiten bei Dokumenten, deren Maßzahl größer als 25% ist. Die Übereinstimmung von 92% im vorliegenden Beispiel rührt von Templates her für ein Kündigungsschreiben (Dokument 0-302) bzw. für eine Kündigungsbestätigung (Dokument 0-299). Bei 100% kann man davon ausgehen, daß es sich praktisch um Doubletten handelt.

Die Suchwörter im Eingabefeld müssen nicht unbedingt in den gefundenen "ähnlichen Dokumenten" vorkommen. Sie bleiben jedoch stehen, damit man einfach (z.B. durch Klick auf *context* oder *treffer*) in die vorherige Recherchephase zurückspringen kann.

(5) Ansichtsoptionen von Dokumenten

Es gibt drei verschiedene Möglichkeiten, sich den Inhalt eines gefundenen Dokuments anzeigen zu lassen.

(i) **Originaldokument.**
Der *treffer*-Bildschirm weist den Pfad mit dem Dateinamen aus. Ein Klick auf diesen Link öffnet das Originaldokument.

(ii) **Dokument im HTML-Format mit Highlight-Funktion**

Durch Klick auf das Symbol  (2. von rechts) öffnet sich eine html-Version des Textinhalts. Der oder die Suchbegriffe sind im Kopf notiert und im Text farblich hervorgehoben (Highlight-Funktion). Durch Scrollen findet man so alle Trefferstellen auf. Alle im Original vorhandenen Links, Grafiken, Bilder etc. sind hier auch vorhanden, können jedoch in der Formatierung und im Layout abweichen.

(iii) **Wie (ii) aber ohne Bilder und Links, dafür mit Sprung zum jeweils nächsten Suchwort.**

Durch Klick auf das Symbol  (ganz rechts) öffnet sich eine Html-Version des Textinhalts. Der oder die Suchbegriffe sind im Kopf notiert und im Text farblich hervorgehoben (Highlight-Funktion). Durch Anklicken des Pfeils in der Kopfzeile bzw. des jeweils "letzten" sichtbaren Trefferworts springt man zum jeweils nächsten. Insofern spart man sich das Scrollen, was besonders bei großen Dokumenten hinderlich ist. Um Konflikte mit bereits bestehenden Links zu verhindern, sind diese aufgehoben.

4 Anwendungsmöglichkeiten und Einsatzfelder

In erster Linie sind die Funktionen der SENTRAX-Engine zum besseren Suchen und Finden in unbekannten, unstrukturierten Textsammlungen konzipiert worden. Die verschiedenen verfügbaren Ähnlichkeitsmaße erlauben aber noch weitergehende Anwendungen. Wenn man sich vorstellt, einen Container mit einer gut definierten Sorte

Texte gefüllt zu haben, z.B. "Kochrezepte" oder "SPAM", dann ließe sich bei einem ankommenden unbekannten Dokument mit den SENTRAX-Funktionen abschätzen, ob es die Merkmale der einen oder der anderen Sorte trägt. So wäre eine automatische Klassifizierung möglich. Im Zusamenhang mit dem Lernen und Arbeiten in netzbasierter Umgebung und mit den Anforderungen des Wissensmanagements ergeben sich wiederum ganz andere Arten der Verwendung dieser Engine. Da die zugrunde liegende Technologie der "Essence Extraction" nicht nur für deutsche Texte sondern auch für viele andere Sprachen wirksam ist, ergeben sich weitere Einsatzmöglichkeiten im Bereich der bi- bzw. multilingualen Recherche. Einige Aspekte dazu sollen kurz angerissen werden.

Wissensmanagement. Sobald ein Unternehmen oder eine Organisation genügend viele Dokumente verfasst und gesammelt hat, gewinnt die Aufgabe, sie zu pflegen, zu sortieren, zu distribuieren, zu archivieren usf. an Bedeutung. Man hat völlig unterschiedliche Sammlungen, wie z.B. Materialien zu Forschungsprojekten, Genehmigungs- und Aufsichtsverfahren, Protokolle, Richtlinien und Gesetzesvorschriften, Handbücher, Präsentationsunterlagen, Archivobjekte. Bei diesen Aufgaben mit textartigen Inhalten kann die SENTRAX-Engine helfen, da unterschiedliche Container definierbar sind, in denen einzeln oder über Auswahlen, aber auch insgesamt gesucht werden kann. Die diversen Funktionen stehen dabei für die jeweiligen Suchanfragen flexibel zur Verfügung, unter Umständen mit spezifischen Parametereinstellungen –z.B. für die Erstellung von Übersichten oder Statistiken. Alle Ausgaben lassen sich auch als Liste darstellen und so wie gewohnt auch automatisch weiterverarbeiten und exportieren.

Kategorisierung. Für die automatische bzw. halbautomatische Kategorisierung von Dokumenten wird in der Literatur gerne die sogenannte SVM-Technik (Support Vector Machine) diskutiert (Kindermann und Leopold [2000]). Die kombinierte Nutzung der SENTRAX Suchfunktionen erlaubt eine alternative Lösungsvariante dieser Problemstellung. Untersuchungen dazu und der Entwurf einer lauffähigen Software wurden erfolgreich auf verschiedene Aufgabenstellungen angewandt. (Vgl. Müller [2002] und Frobese [2006]).

Aus- und Weiterbildung. Definierte Ontologien müssen gut gepflegt werden, insbesondere wenn sie über einen längeren Zeitraum (=mehrere Generationen von Nachwuchsmitarbeitern) gültig, zutreffend, verwendbar sein sollen. Die von der Engine erzeugten *context*-Begriffswolken zu einem Thema oder zu Themenzusammenhängen unterstützen diese Erstellungs- und Pflegearbeiten.

Erzeugt man aus einer geeigneten Dokumentensammlung die zu ausgewählten Fachbegriffen gehörende *context*-Begriffswolke, so kann man diese als „Test“ verwenden, indem man die Probanden zur Bedeutung der Begriffe befragt bzw. deren eventuelle Zusammengehörigkeit erläutern läßt. Ist die zu testende Person bei dem einen oder anderen Fachbegriff unsicher, genügt häufig ein Klick darauf, um mit dem neuen Screen weitere Informationen bzw. Hilfen zu sehen. Gibt man z.B. "statistische Textanalyse" (bei einem passenden Container) ein, dann liefert die SENTRAX die Wortbeziehungen in der Abbildung 6. Es sind mehrere „Fragenkomplexe“ zu sehen, also Begriffsgruppen, die auf Themen referieren, welche in der Ausbildung von Interesse waren. Zum Beispiel: „Was hat statistische Textanalyse mit Rechtschreibfehlerkorrektur zu tun?“ „Was versteht man unter Trigramm

und welche Rolle spielt das (hier)?“ „Welchen Zusammenhang kann Schokolade mit Stimuluswörter haben?“ Usw.

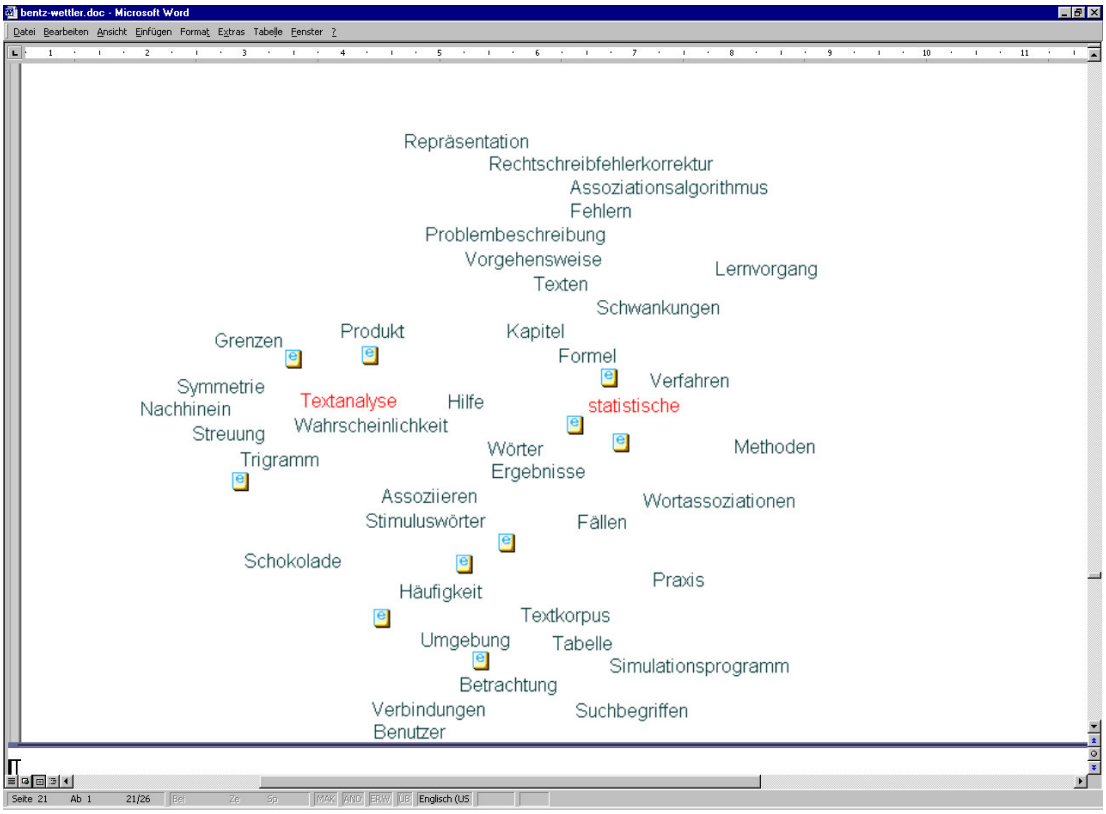

Abb. 6: SENTRAX *context*-Screen zu "statistische Textanalyse", vgl. Bentz [2006]

Wenn man solche Prüfungsscreens nicht aus Forschungsbeiträgen erzeugt, sondern vielmehr eine Sammlung von Lehrbuchtexten zugrunde legt, dann sollten die Gruppierungen noch viel prägnanter werden. Als ein Beispiel dafür wurde eine WBT (Web Based Training)-Einheit zum Thema „Fettstoffwechsel“ indexiert und diese den Teilnehmern der Lernplattform ApoLearn in den Zugriff gestellt. (Vgl. www.apolearn.de bzw. den Link auf der Seite www.imbyte.de). Hier kann der Lernende den Context-Bildschirm zur Orientierung nutzen und von da aus direkt auf das Kapitel zugreifen, in dem das Thema von den Eingabewörtern bzw. den Antwortclustern handelt.

Erfassung von Expertenwissen. Die Ergebnisse der SENTRAX-Contextmap können nicht nur als Antwort auf eine Suchanfrage, sondern als Darstellung eines gewissen Umfeldes zum Thema der eingegebenen Suchbegriffe interpretiert werden. Insofern fungiert die Engine wie ein „passives Gedächtnis“, fördert also sehr effizient die Entscheidung, welche Art von Wissensdomänen oder Wissensaussagen mit ihren charakteristischen Fachbegriffen notiert werden können bzw. müssen. Diese Art Gedächtnis wird aber eine weitaus größere Menge Daten erfassen können als der Mensch es vermag und dabei nie etwas vergessen. Wenn man die Parameter gut einstellt, kann schnell und zuverlässig eine komplette Übersicht aller Vorkommen und Verbindungen von „Konzepten“ in den Dokumenten erhalten werden. Für diese Aufgabe und Anwendung wurde eine Hilfsfunktion *mindshift* implementiert, die die intellektuelle Bearbeitung und Speicherung von Ergebnisbildschirmen erlaubt.

Sprachenübergreifendes Information Retrieval. Im Zusammenhang mit bilingualen Retrievalaufgaben kann man mit Hilfe der SENTRAX-Technologie die Hypothese untersuchen, ob parallele Textkorpora „ähnliche“ Clusterformationen aus den beiden den Sprachen zugeordenten Containern hervorbringen. Man wird dabei zunächst voraussetzen, daß die beiden Sprachen „ähnliche“ Regeln haben, wie etwa Englisch-Deutsch. Die Frage erlaubt eine positive Antwort, wie die Untersuchungen von Suriya Na nhongkai zeigen [2006]. Mit diesem Ansatz kann man sogar inhaltliche Aussagen über Dokumente machen, ohne deren Übersetzung zu haben. Die Ursache liegt darin, daß die Maschine andere Merkmale generiert und verarbeitet als der Mensch, wenn es darum geht, mehrsprachige Dokumente aus einer größeren Sammlung einander zuzuordnen.

Abb. 7: Die SENTRAX *context*-Screens zweier paraller Dokumente (D-E)

5 Die Ähnlichkeitsmaße der SENTRAX Engine

Während die *lexico*-Funktion String-orientiert arbeitet, hauptsächlich auf der Basis von n-Grammen, findet die *context*-Funktion bedeutungsverwandte Begriffe in den Dokumenten. Dazu werden Auftretenshäufigkeiten und nahes Beieinanderstehen von Worten und Wortgruppen in den Texten ausgewertet. Man hat daher oft semantisch verwandte Begriffe in der *Context*Map, wie z.B. *Fusion-Zusammenschluss*, es werden aber auch gänzlich verschiedene Worte dort zusammengebracht, wie z.B. *Ausbildung-Analphabetentum*, weil sie durch die Art ihres Auftretens in den Dokumenten einen Vorgang oder eine Idee repräsentieren. Die Güte dieser Funktion hängt von der Homogenität des Datenmaterials ab. Für normale Texte, die aus ordentlichen Sätzen bestehen, funktioniert die *Context*Map ziemlich gut. Für Wörter, die inhaltlich zusammenhangslos in Tabellen stehen, wie z.B. in Telefonlisten, darf nicht zuviel von dieser Funktion erwartet werden, da der „Kontext“ vom Benutzer nicht zuverlässig interpretiert werden kann.

Die *treffer*-Funktion zeigt alle Dokumente, in denen die Suchwörter enthalten sind mit 100%-Güte an. Im Falle des Fehlens eines oder mehrerer Sucheingaben wird die Ausgabeliste entsprechend modifiziert, so dass ein betroffenes eine Rangabstufung erfährt. Dokumente auf gleicher Stufe werden nach ihrer intern vergebenen ID sortiert. Innerhalb einer festen Prozentgruppe sind also alle Dokumente gleich gut.

Die *similarDoc*-Funktion arbeitet wieder auf den Wörtern (jetzt des gesamten Textes) und sucht entsprechend passende Dokumente zusammen. Auch hier sind alle Treffer auf derselben Prozentstufe gleichermaßen gut. Diese Funktion ist nicht notwendig symmetrisch, was das Empfinden des Benutzers normalerweise nicht stört. Denn auch ohne IR-Systeme kann es vorkommen, daß ein Dokument A bestpassend zum Dokument B ist, B wiederum (weil es vielleicht viel umfangreicher als A ist) besser zu C passt usw.

6 Fazit

Indem sich die SENTRAX von den herkömmlichen Matching Algorithmen und invertierten Listen zur Indexierung löst und eine aufgabenbezogene Mustererkennung mit angepassten Ähnlichkeitsmaßen verwendet, ermöglicht sie eine fehlertolerante und flexible Suche im Datenbestand. Die Visualisierungen über die *Lexico*Map und die *Context*Map bieten eine wertvolle Hilfe beim Erforschen des Korpus und Verfeinern der Suche. Weit über die "normale" Wortsuche hinaus wird in dem neuen Ansatz vor allem das im Information Retrieval vorherrschende Problem der unterschiedlichen Konzeptrepräsentation in Angriff genommen. Verschiedene Autoren haben unterschiedliche Wortwahlen, um bestimmte Vorgänge, Vorfälle, Ideen oder Konzepte zu beschreiben. Die Suchenden wiederum bedienen sich oft noch anderer Begriffe, um Informationen zu diesen Themen aufzufinden. Bei den Suchenden kommt hinzu, dass sie häufig ein „Informationsbedürfnis“ (information need) zu stillen versuchen und vielleicht zunächst gar keine klare Vorstellung davon haben, wie und mit welchen Begriffen sie am besten ans Ziel kommen.

Da nun die SENTRAX Engine die Suchbegriffe innerhalb von Wortwolken visualisiert und ähnliche, d.h. häufig kookkurrierende Begriffe anzeigt, wird dem Nutzer so ein Bild vermittelt, welche Begriffe in seinem Suchfeld eine wichtige Rolle spielen. Er hat damit die Möglichkeit, die Suche schrittweise zu verfeinern und zu präzisieren („Query Reformulation“). Die SENTRAX unterstützt ihn also zielgerichtet beim Auffinden begriffsverwandter Begriffe, die bei invertierten Listen entschlüpfen. (Vgl. Kummer [2006])

Man kann zusammenfassend sagen, daß dem Nutzer mit dieser Technologie ein schnellerer Zugang zu notwendigen Informationen ermöglicht und ein effizienterer Zugriff auf relevante Datenobjekte als bei herkömmlichen Ansätzen gewährt wird.

Literaturverzeichnis

Ackermann, Martin (2000): Statistische Korpusanalyse zum Extrahieren von semantischen Wortrelationen. Dissertation. Hildesheimer Informatik-Berichte 1/2000. Hildesheim: Universität Hildesheim.

Bentz, Hans-Joachim (2002): Lernen und Arbeiten in virtuellen Räumen - Bezüge zu Wissensmanagement, E-HRM & E-Business. In: Handbuch E-Learning: Expertenwissen aus Wissenschaft und Praxis. Hohenstein, Andreas; Wilbers, Karl (Hrsg.). Köln: Deutscher Wirtschaftsdienst.

Bentz, Hans-Joachim (2006): Suchen und Problemlösen in komplexer Umgebung. In: Perspectives on Cognition: A Festschrift for Manfred Wettler. Rapp, Reinhard; Sedlmeier, Peter (Hrsg.). Lengerich: Pabst Science Publishers.

Frobese, Dirk (2006): Suchmethoden der KI angewendet auf elektronische Nachrichten (E-Mails). Universität Hildesheim. Unveröff. Manuskript.

Kindermann, J.; Leopold, E. (2000): Classification of Texts with Support Vector Machines. An Examination of the Efficiency of Kernels and Data-Transformations; 24th Annual Conference of the Gesellschaft für Klassifikation; Passau.

Kummer, Nina (2006): Analyse von Trefferlisten herkömmlicher Suchmaschinen. Universität Hildesheim, Masterarbeit.

Müller, Karen (2002): Automatische Klassifikation von Textdokumenten. Universität Hildesheim. Masterarbeit.

Na nhongkai, Suriya (2006): Untersuchungen zur sprachübergreifenden, bilingualen Suche mit Hilfe der Konzeptnetz-Technologie der SENTRAX-Engine. Universität Hildesheim. Dissertation.

Wettler, M.; Ferber, R.; Rapp, R. (1995). An associative model of word selection in the generation of search queries. Journal of the American Society for Information Science,46 (1995), 685-699.